GEAR-Align: Grounding-Evidence-Aware Gradient Routing for Multimodal Alignment

Yongkang Yu[1*] Haobo Wang[2*] Meng Chen[3] Han Fang[4] Xin Wei[4] Zhiyu Lin[4] Ye Yuan[4] Chiming Duan[2] Hao Sun[4†] Haiyang Zhang[1†]

[1] Beijing University of Posts and Telecommunications [2] Peking University [3] Fudan University [4] Institute of Artificial Intelligence (TeleAI), China Telecom

NEURAL INFORMATION PROCESSING SYSTEMS 2026

Motivation

Not every token needs visual evidence.

Standard SFT can reduce prediction loss without strengthening dependence on the correct image.

L Language-predictable G General-grounding E Evidence-specific

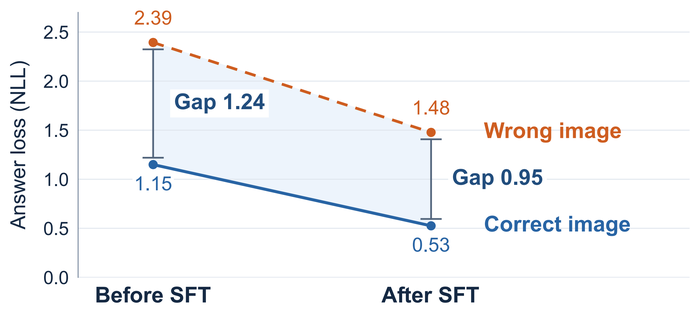


Stop sign

The image is dominated by a red octagonal stop sign positioned in the center. The word STOP is printed in large white capital letters, and the sign is mounted on a gray metal pole.

Representative G: positioned, mounted
Representative E: red, stop, printed, large, gray

Lower loss, weaker visual dependence



Better answer fitting
1.15 → 0.53
Loss with the correct image

Smaller image advantage
1.24 → 0.95
Wrong − correct image loss

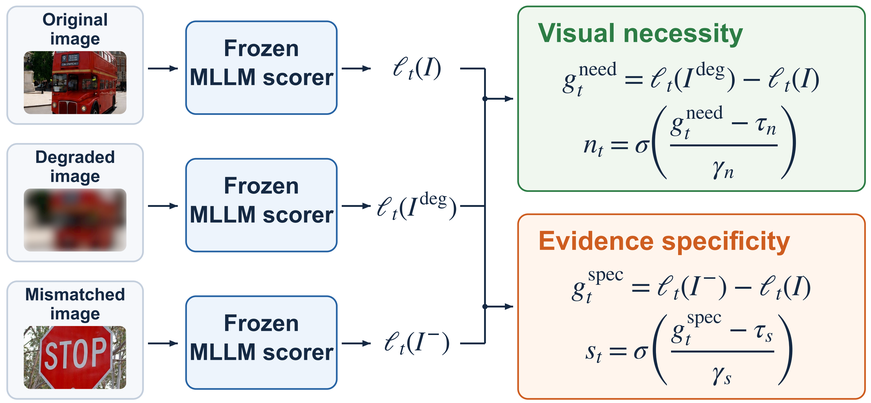
Degraded images show the same trend: gap 1.02 → 0.74.
Evidence-dependent token subset · Standard SFT

Takeaway
Answer fitting alone does not ensure visual grounding.

GEAR-Align

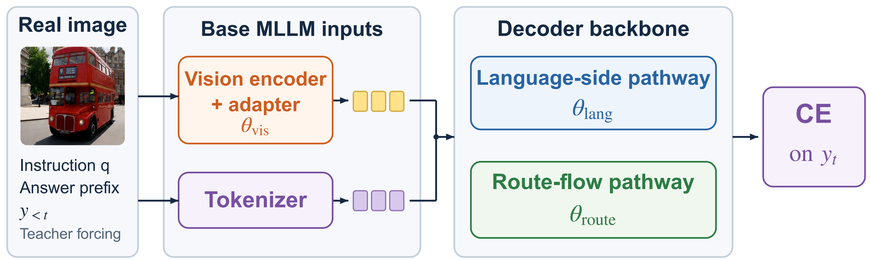
1. Offline counterfactual profiling

Compare token NLLs from a frozen scorer.



Saved scores n_t and s_t are used only to weight training losses.

2. Online gradient routing



θ_{route} includes designated decoder layers + inserted modules.

Route-flow module Image + instruction prefix → visual-token update

$$H_v, H_q \longrightarrow R(H_v, H_q) \longrightarrow H_v \leftarrow H_v + R(H_v, H_q)$$

Shared prefix computation in training and generation.
Cumulative weights; each loss updates its own parameter group.

Language	Route-flow	Visual-evidence
$\rho_t^{\text{lang}} = 1$	$\rho_t^{\text{route}} = n_t$	$\rho_t^{\text{vis}} = n_t s_t$

Objective $\mathcal{L} = \mathcal{L}_{\text{lang}} + \lambda_{\text{route}}\mathcal{L}_{\text{route}} + \lambda_{\text{vis}}\mathcal{L}_{\text{vis}}$

Experimental evidence

Main results across backbones

Backbone	Method	HallusionBench ↑	OCRBench ↑	TextVQA ↑
Qwen3-8B	SFT	38.9	73.4	77.4
	GEAR-Align	**43.3**	**77.5**	**79.5**
Qwen3-4B	SFT	32.1	63.8	68.0
	GEAR-Align	**38.0**	**69.3**	**71.5**
LLaVA-NeXT 8B	SFT	26.0	70.0	70.8
	GEAR-Align	**29.1**	**73.2**	**73.3**

Selected benchmarks; full comparisons in the paper.

Key ablations on Qwen3-8B

Setting	HallusionBench ↑	OCRBench ↑
Score-weighted SFT	42.1	76.8
Joint parameter updates	39.6	73.8
Shuffled token profiles	42.0	75.8
Necessity-only supervision	40.2	75.9
GEAR-Align (full)	**43.3**	**77.5**

The full policy outperforms scalar reweighting and joint parameter updates on these benchmarks.

Compute control on Qwen3-8B

Method	Samples	Total GPU-h	HallusionBench ↑
SFT	300K	92.0	38.9
GEAR-Align	**192K**	**91.5**	**42.2**

Approximate total includes profiling and online training.

Conclusion
Match gradient destinations to visual dependence.

* Equal contribution † Corresponding authors